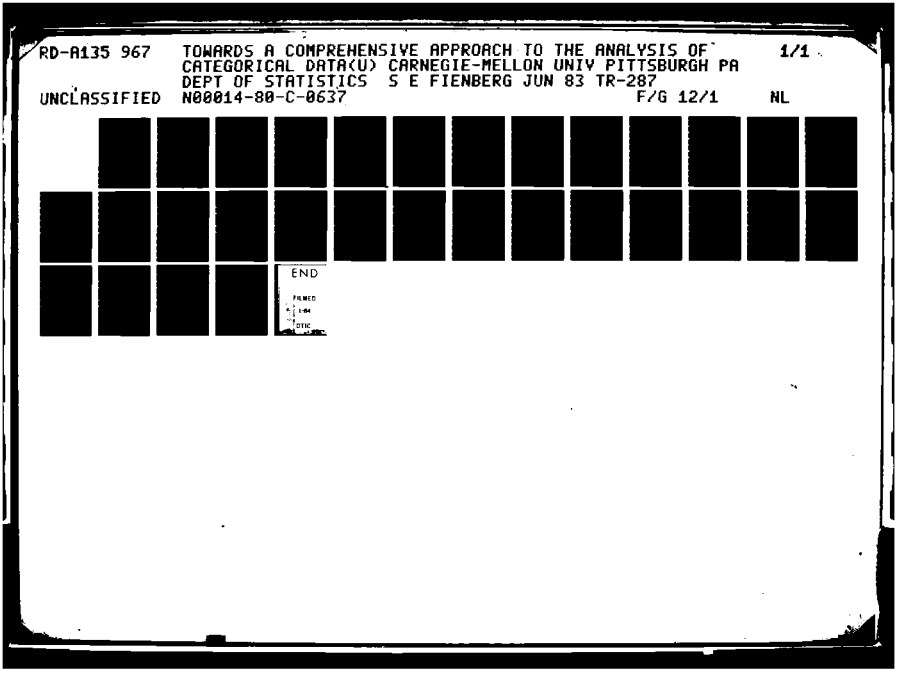


RD-A135 967 TOWARDS A COMPREHENSIVE APPROACH TO THE ANALYSIS OF CATEGORICAL DATA(U) CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS S E FIENBERG JUN 83 TR-287 1/1

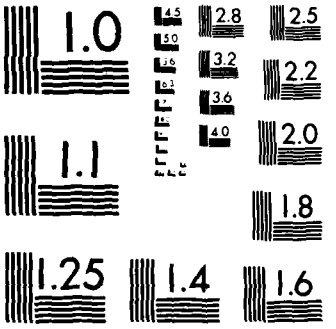
UNCLASSIFIED N00014-80-C-0637 F/G 12/1 NL

END

FILMED

1-84

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A135 967

(1)

TOWARDS A COMPREHENSIVE APPROACH TO THE ANALYSIS OF CATEGORICAL DATA*

by

Stephen E. Fienberg

Technical Report No. 287

Department of Statistics

Carnegie-Mellon University

Pittsburgh, PA 15213

June, 1983

* To be presented at *Statistics: An Appraisal*, an International Conference to Mark the 50th Anniversary of the Iowa State University Statistical Laboratory, June 13-15, 1983. The preparation of the paper was supported in part by Contract N00014-80-C-0637 from the Office of Naval Research to Carnegie-Mellon University. Reproduction in whole or part is permitted for any purpose of the United States Government.

DTIC FILE COPY

This document has been approved for public release and sale; its distribution is unlimited.

DEC 1 1983

88 12 12 014

SUMMARY

Current methodology for the analysis of categorical data can be traced to two papers published in 1900, one by Pearson and one by Yule. The keys to the linking of their ideas have been the use of linear models and inferential tools due to Fisher. After 50 years of research, statisticians have come close to developing a comprehensive approach to categorical data problems that stresses three basic themes: interpretability, flexibility, and computability. This paper surveys the evolution of this comprehensive approach and classes of problems for which it has proven useful. A concluding section contains some speculation on unsolved methodological problems of current interest and on future developments.

Key Words: Chi-square goodness-of-fit tests; Conditional independence; Cross-classifications; Loglinear models; Interactions; Rectangular arrays.

DTIC COPY INSPECTED

A-1

1. Introduction

As recently as the late 1960's, the perception of most statistics students and users of statistical methods was that the analysis of categorical data consisted primarily of topics such as 2×2 tables and Fisher's exact test, chi-square tests for goodness-of-fit, and examining the models of independence or homogeneity of proportions in two-way contingency tables. The reality was that by 1965 the statistical literature contained over 150 papers on methodology for contingency table analysis including techniques for the analysis of multi-way tables using loglinear models (e.g. see the partial bibliography in Kastenbaum, 1970, which goes up to 1965, and the subsequent bibliography through 1974 by Killion and Zahn, 1976). The intervening 15 years have seen a dramatic change in both the perception and the reality. The development and elaboration of the loglinear model approach to categorical data analysis has led to the publication of at least a dozen books and monographs on the topic (for a partial list see Fienberg, 1982a), and the basic ideas on the use of loglinear models for multi-way arrays now appear in many textbooks on statistical methodology beside material on multiple regression and ANOVA, whose linear heritage they share.

The key features of what is viewed by many as a comprehensive approach to the analysis of categorical data can be traced back to two unrelated papers, both of which appeared in 1900. In one of these papers, Pearson (1900) proposed the chi-square test for comparing observed and expected frequencies, and derived its asymptotic χ^2 distribution when the parameters underlying the expected frequencies are known a priori (for further details and a discussion, see Plackett, 1983). This result, as amplified by Fisher (1922) to adjust the degrees of freedom (d.f.) for the estimation of parameters, forms the basis of the usual asymptotic theory used to check on the goodness-of-fit of loglinear and other models. In the other paper, Yule (1900) described the structural relationship among categorical variables by means of functions of cross-product or odds ratios. In particular he developed a general notation for 2^n contingency tables and the concepts of partial and joint association for dichotomous variables. Fisher (1922) did pull these ideas together for I×J contingency tables, showing that the chi-square test

for independence had $(I-1)(J-1)$ d.f. This is roughly where the development of categorical data analysis stood when Fisher first visited Iowa State University in the summer of 1931, just before the founding of the Statistical Laboratory. (It was on this occasion that Fisher learned that A.E. Brandt had developed a formula for computing the chi-square statistic in a special case, and Fisher incorporated into one of his lectures at Ames (Box, 1978).)

Over the years, there has been considerable interest in the analysis of categorical data at Iowa State University. Snedecor (1937) included material on it in the first edition of *Statistical Methods*, and Cochran (1940, 1942) wrote on the topic during his ISU years. Later editions of *Statistical Methods* incorporated Bartlett's work on 2×2×2 tables although Snedecor's (1958) paper incorrectly noted that Bartlett's test for no-second-order interaction and a test proposed by Lancaster are "asymptotically equal." Other ISU faculty and graduates who have made methodological contributions to the topic include R.L. Anderson, K. Hinkelman, O. Kempthorne, and K. Koehler.

The next section of this paper gives a brief historical review of the development of ideas on loglinear models and their use in the analysis of categorical data over the past 50 years. Then in Section 3, we describe the use of loglinear models for contingency tables, stressing alternate representations of the models and their interpretations. In Section 4 we indicate how loglinear models have been adapted to other forms of categorical data analysis, and the links between these new methods and loglinear models for multi-way contingency tables that facilitate computation of parameter estimation. We conclude the paper with some speculation on unsolved methodological problems of current interest and on future developments.

No single approach can ever be expected to be the only sensible one for a broad class of statistical problems such as those associated with the analysis of categorical data. Yet the interpretability and flexibility of the loglinear model approach and the computational methods available for its application have moved us towards a comprehensive approach to the analysis of categorical data.

2. A Brief Review of Loglinear Model Developments

The literature on the analysis of categorical data contains hundreds of papers authored by many of statistics' most distinguished researchers. In this section, we trace a path through this literature of the past 50 years that highlights the evolution of the loglinear model and its application. This brief review ignores the contributions of a large number of individuals who focussed primarily on other forms of models, methods of estimation other than maximum likelihood, and issues such as the adequacy of large-sample properties of test statistics. For an alternative review and a discussion of nonstandard applications see Imrey, Koch, Stokes, et al. (1981, 1982).

Although Yule (1900) focussed on the cross-product ratio as a measure of association in 2×2 tables and developed ideas on association in 2^n tables, 35 years passed before Bartlett (1935) utilized Yule's ideas to define the concept of second-order interaction in 2×2×2 tables. For a 2×2 table with expected values $\{m_{ij}\}$, Yule's cross-product ratio is:

$$\alpha = \frac{m_{11}m_{22}}{m_{12}m_{21}} \tag{2.1}$$

Bartlett's no-second-order interaction model for the expected values in a 2×2×2 table

m_{111}	m_{121}	m_{112}	m_{122}
m_{211}	m_{221}	m_{212}	m_{222}

was based on equating the values of α in each layer of the table, i.e.,

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}} \tag{2.2}$$

Bartlett then went on to derive maximum likelihood estimates of the $\{m_{ijk}\}$ by solving a cubic equation.

It was not for another 20 years that Roy and Kastenbaum (1956) were to generalize Bartlett's approach to I×J×K tables. Their method, as described in Kastenbaum and Lamphiear (1959), for solving the likelihood equations under no-second-order interaction was considered to be

computationally complex (involving an iterative solution of (I-1)(J-1)(K-1) simultaneous third-degree equations), and neither the model nor the method was easily generalized to higher dimensions. Indeed, it was not until Birch (1963) converted Bartlett's and Roy and Kastenbaum's multiplicative definition of no-second-order interaction to an additive analysis-of-variance-like model in the logarithmic scale that key features of the loglinear model approach to multi-way tables emerged (see also the development in Good, 1963). Birch also presented a simple yet elegant result that linked the basic sampling distributions for contingency tables (Poisson and multinomial) and at the same time elucidated the relationship between loglinear and logit models. What remained to be done before the approach could be implemented in practice was to come up with a simple computational technique for solving the likelihood equations.

The timing was propitious because iterative techniques that involved large numbers of computations had recently become a reasonable way to solve maximization problems due to the availability of high-speed computers. While working on the National Halothane Study in 1965-66, Bishop rediscovered an iterative procedure proposed for a related categorical data problem by Deming and Stephan (1940). Although others (e.g. see Darroch, 1962) had proposed equivalent iterative techniques for special cases, Bishop (1967) presented a relatively general computer program implementing the Deming-Stephan algorithm and showed how it was applicable for solving the likelihood equations associated with the class of loglinear models described by Birch.

Many statisticians were now focussing on loglinear model methods, and adapting them for use in connection with the analysis of incomplete contingency tables, Markov chains, and other non-standard problems. Important advances were made by authors such as Bhapkar, Bock, Darroch, Goodman, Haberman, Kullback, Plackett, and Nerlove and Press. One specific line of work, initiated by Nelder and Wedderburn (1972), linked the analysis of categorical data using loglinear and logit models to the analysis of measurement data linear models with normal errors via what they called generalized linear models. As implemented in the computer

package GLIM (Baker and Nelder, 1978), this approach provided additional stimulus for the use of loglinear models and presented an alternative to the iterative proportional fitting technique introduced by Bishop.

The research work of the 1960's treated the problems associated with categorical data analysis using loglinear models as being separate from those involving other forms of linear models and sampling distributions other than Poisson and multinomial. But as the work of Nelder and Wedderburn showed, these separate streams of research could be linked. The key to the linkage was the existence of general results on exponential families and their sufficient statistics that originated in the 1930's with Fisher (e.g. see Dempster, 1971, and the discussion in Andersen, 1980). From the perspective of exponential family theory the interpretation of loglinear models was even closer to that of linear models than the parallel notation suggested. In the next section, we describe some of the loglinear model results that are part of this more general statistical theory, but we also stress special aspects of the interpretation of loglinear models and a unique loglinear/multinomial result due to Birch.

3. Loglinear Models, Contingency Tables, and Likelihood Theory

A. Notation for the 2×2 table

It has been suggested, only partially in jest, that virtually all important statistical ideas can be described and illustrated in the context of the 2×2 contingency table. While this is clearly not the case, the 2×2 table provides a useful starting place for a discussion of loglinear models.

We begin by denoting the observed count for the (i,j) cell of a 2×2 contingency table by x_{ij} and the totals for the ith row and jth column by x_{i+} and x_{+j}, respectively. The $\{x_{ij}\}$ are typically taken to be realizations of random variables whose expectations we denote by $\{m_{ij}\}$. These expected values can now be rewritten in loglinear model form using analysis of variance (ANOVA) notation:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \tag{3.1}$$

where

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0. \tag{3.2}$$

Although the model is in a general form to be applicable to I×J tables, for 2×2 tables there are only 4 distinct parameters: u, $u_{1(1)}$, $u_{2(1)}$, and $u_{12(11)}$. The 3 subscripted parameters are expressible as

$$u_{12(11)} = \tfrac{1}{4} \log \frac{m_{11}m_{22}}{m_{12}m_{21}}, \tag{3.3}$$

$$u_{1(1)} = \tfrac{1}{4} \log \frac{m_{11}m_{12}}{m_{21}m_{22}}, \tag{3.4}$$

and

$$u_{2(1)} = \tfrac{1}{4} \log \frac{m_{11}m_{21}}{m_{12}m_{22}}. \tag{3.5}$$

We note that $u_{12(11)}$ is simply a function of Yule's cross-product ratio, $\alpha = m_{11}m_{22}/m_{12}m_{21}$, and $u_{1(1)}$ and $u_{2(1)}$ are functions of similar cross-product ratios.

Setting $u_{12(11)} = 0$ is equivalent to setting $\alpha = 1$ and corresponds to independence of the variable for rows and the variable for columns. Thus we have seen two special features of loglinear models:

(i) all subscripted parameters are expressible as logarithms of cross-product ratios or functions of them,

(ii) setting some loglinear model parameters equal to zero often leads to a model which can be interpreted in terms of independence of variables underlying the dimensions of the table.

These features, which are shared by loglinear models for I×J and multi-way tables, mean that loglinear models can be interpreted using both the ANOVA-like structure or generalizations of cross-product ratios and independence concepts.

The use of ANOVA-like notation here is at least in part illusory, however. There is no response variable on the left-hand side of equation (3.1), only a log-expected count. Thus the u-term parameters really cannot be thought of as "effects" of one variable on another. This form of ANOVA interpretation will prove useful only when we can convert a loglinear model

into a logit model, as we illustrate in the next subsection.

B. Loglinear models for I×J×K tables

For a three-way table of counts, $\{x_{ijk}\}$, the general loglinear model for the corresponding expected values, $\{m_{ijk}\}$, can be written as:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \tag{3.6}$$

where, as in the usual ANOVA model, all subscripted parameters sum to zero over each subscript, e.g.

$$\sum_i u_{1(i)} = \sum_i u_{12(ij)} = \sum_i u_{123(ijk)} = 0. \tag{3.7}$$

In the special case where I = J = K = 2, there are only 8 distinct parameters: u, $u_{1(1)}$, $u_{2(1)}$, $u_{3(1)}$, $u_{12(11)}$, $u_{13(11)}$, $u_{23(11)}$, and $u_{123(111)}$. Each of the 7 subscripted parameters are expressible as a function of the ratio of two cross-product ratios, e.g.

$$u_{123(111)} = \frac{1}{8} \log \left(\frac{m_{111} m_{221}}{m_{121} m_{211}} \Big/ \frac{m_{112} m_{222}}{m_{122} m_{212}} \right) \tag{3.8}$$

and

$$u_{13(11)} = \frac{1}{8} \log \left(\frac{m_{111} m_{121}}{m_{211} m_{221}} \Big/ \frac{m_{112} m_{122}}{m_{212} m_{222}} \right). \tag{3.9}$$

These expressions are standard ANOVA-like contrasts for the log-expected counts. In an I×J×K table each subscripted u-term can be rewritten as a linear combination of the logarithm of the ratio of cross-product ratios associated with the corresponding parameters for all possible 2×2×2 subtables.

In the 2×2×2 table, setting $u_{123(111)} = 0$ is equivalent to Bartlett's condition for no-second-order interaction given in expression (2.2). In the I×J×K table, setting $u_{123(ijk)} = 0$ for all i, j, and k is equivalent to Roy and Kastenbaum's generalization of Bartlett's condition. This is one of four special cases of the general loglinear model, (3.6), found by setting sets of u-terms equal to zero:

(a) $u_{123(ijk)} = 0$,

(b) $u_{12(ij)} = u_{123(ijk)} = 0$, (3 versions)

(c) $u_{12(ij)} = u_{13(ik)} = u_{123(ijk)} = 0$, (3 versions)

(d) $u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$,

each for all i, j, and k. The other three special cases each can be re-expressed so that an interpretation in terms of independence is possible. Model (d) corresponds to complete independence among the variables for the three dimensions of the table. Model (c) corresponds to independence of variable 1 and variables 2 and 3, considered jointly. Finally Model (b) corresponds to conditional independence of variables 1 and 2 given the value of variable 3.

Once again we get, in addition to the ANOVA structure, the two features of loglinear models alluded to above:

(i) all subscripted parameters are expressible as logarithms of ratios of cross-product ratios or functions of them,

(ii) several special cases of the general model are interpretable in terms of independence or conditional independence.

To these features we can add a third when there is a distinction between explanatory and response variables for the underlying dimensions.

Let us begin with the case of a 2×J×K table in which the first variable is the response, and the other two are explanatory. The odds of being in category 1 of the response variable versus being in category 2, given the levels of the explanatory variables, are a natural quantity of interest. The log-odds can be expressed in terms of the loglinear model parameters simply by taking differences, i.e.

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = \log m_{1jk} - \log m_{2jk}$$

$$= 2\left[u_{1(1)} + u_{12(1j)} + u_{13(1k)} + u_{123(1jk)}\right].$$

Relabelling the u-terms using a new set of parameters $w = 2u_{1(1)}$, $w_{2(j)} = 2u_{12(1j)}$, $w_{3(k)} = 2u_{13(1k)}$, and $w_{23(jk)} = 2u_{123(1jk)}$ we get the *logit* model:

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = w + w_{2(j)} + w_{3(k)} + w_{23(jk)}, \tag{3.10}$$

where

$$\sum_j w_{2(j)} = \sum_k w_{3(k)} = \sum_j w_{23(jk)} = \sum_k w_{23(jk)} = 0. \tag{3.11}$$

The ANOVA-like parameters in this logit model are interpretable in terms of the "effects" of the explanatory variables on the log-odds of the response. For example, $w_{23(jk)}$ is the interactive effect of variables 2 and 3 on the log-odds when variable 2 is at level j and variable 3 is at level k over and above the separate effects for variable 2 and 3. Note that none of the u-terms in the loglinear model involving only the explanatory variables are present in the logit version of the model.

For an I×J×K table in which the first variable is the response, the loglinear model of expression (3.6) can be rewritten as a set of I-1 logit models for the log-odds,

$$\log\left(\frac{m_{ijk}}{m_{1jk}}\right) \qquad i = 1, 2, \ldots, I-1,$$

with each logit model being of the form of expressions (3.10) and (3.11). If we use a transformation other than logarithmic for the odds in (3.12), then we get other members of Nelder and Wedderburn's GLIM family. For example, the probit or integrated normal scale is

$$\Phi^{-1}\left[m_{ijk}/(m_{ijk} + m_{1jk})\right]$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative normal c.d.f. Among the members of the GLIM family, only the logit (or loglinear) model includes as special cases the models that are interpretable in terms of independence and conditional independence of the underlying variables.

Even in the absence of the statistical estimation results in the following subsection, the interpretability of loglinear models makes them an ideal candidate for the basis of a comprehensive approach to the analysis of categorical data.

C. Key results from likelihood theory

There are three standard sampling models for the observed counts in contingency tables. We

begin by describing them for a singly subscripted vector of t cells, $x^T = (x_1, x_2, ..., x_t)$. This notation for an arbitrarily structured collection of t cells will prove to be of great use in the non-contingency-table problems described in the next section of the paper. For the 2×2 table t = 4, and for the general three-way table t = IJK. Now let $m^T = (m_1, m_2, ..., m_t)$ be the vector of expected values that are assumed to be functions of unknown parameters $\theta^T = (\theta_1, \theta_2, ..., \theta_s)$, where $s < t$. Thus we can write $m = m(\theta)$. The three sampling models are:

POISSON MODEL. The $\{x_i\}$ are observations from independent Poisson random variables with means $\{m_i\}$ and likelihood function

$$\Pi_{i=1}^{t} \left[m_i^{x_i} \exp(-m_i)/x_i! \right]. \tag{3.12}$$

MULTINOMIAL MODEL. The total count $N = \Sigma_{i=1}^{t} x_i$ is a random sample from an infinite population where the underlying cell probabilities are $\{m_i/N\}$, and the likelihood is

$$N! \cdot N^{-N} \Pi_{i=1}^{t} (m_i^{x_i}/x_i!). \tag{3.13}$$

PRODUCT-MULTINOMIAL MODEL. The cells are partitioned into sets, and each set has an independent multinomial structure, as in the multinomial model.

Loglinear models in this setting come about by representing the vector of log expectations $\lambda^T = (\log m_1, ... \log m_t)$ as a linear combination of this parameter in the vector θ. The following pair of results now follow directly from exponential family theory for the Poisson and multinomial sampling schemes.

RESULT 1. Corresponding to each parameter in θ is a minimal sufficient statistics (MSS) that is expressible as a linear combination of the $\{x_i\}$. (More formally, if M is used to denote the loglinear model specified by $m = m(\theta)$, then the MSS's are given by the projection of x onto M, $P_M x$.

RESULT 2. The maximum likelihood estimate (MLE), $\hat{m}$, of m, if it exists, is

unique and satisfies the likelihood equations

$$P_M\hat{\mathbf{m}} = P_M\mathbf{x}. \tag{3.14}$$

Necessary and sufficient conditions for the existence of a solution to the likelihood equations of expression (3.14) are given in Haberman (1974), and for various special cases by a variety of authors. Nonexistence occurs when the likelihood is maximized on the boundary of the parameter space, and this corresponds to some $\hat{m}_i$'s being equal to zero. Although deriving constructive conditions for the existence of MLE's has been viewed by many as an esoteric research problem, in fact it is a critical component to computational methods for solving the likelihood equations and is of great practical import for those who wish to analyze large sparse contingency tables.

We now come to the third key result, which was first given by Birch (1963) and which unifies the three sampling schemes and links the MLE's for loglinear and logit models. For product-multinomial sampling situations, the basic multinomial constraints (i.e., that the counts must add up to the multinomial sample sizes) must be taken into account. One way to think about this in the context of loglinear models is to recall that, from Result 1, these sample sizes are marginal totals, which under a simple multinomial or Poisson model are MSS's corresponding to some of the parameters in θ specifying the loglinear model M, i.e., $\mathbf{m} = \mathbf{m}(\theta)$ are fixed by these constraints. What we do is consider a logit model, M^*, where these components of θ "drop out."

More formally, let M^* be a logit model for $\mathbf{m}$ under product-multinomial sampling which corresponds to a loglinear model M under Poisson sampling such that the multinomial constraints "fix" a subset of the parameters, θ, used to specify M. Then Birch's result is:

RESULT 3. The MLE of $\mathbf{m}$ under product-multinomial sampling for the model M^* is the same as the MLE of $\mathbf{m}$ under Poisson sampling for the model M.

This result is directly related to a more general theorem from exponential family theory

which states that, if we have an exponential family density in minimal form with MSS's h_1, h_2, ..., h_s, then the conditional density for h_1, h_2, ..., h_k given h_{k+1}, h_{k+2}, ..., h_s has the same exponential family form. Moreover, the exponential family parameters for this conditional density are the ones from the original density for which h_1, h_2, ..., h_k are MSS's, and h_1, h_2, ..., h_k are the corresponding MSS's in the conditional density (see Andersen, 1980, pp.82-83 for a formal statement and proof). What is so special about Result 3, the loglinear/multinomial version of this theorem especially in the context of contingency tables, is that *the MSS's, $P_M\mathbf{x}$, are marginal totals for the original vector, $\mathbf{x}$, and thus the conditional density has the minimal form for the conditional distribution of the response variables given a set of explanatory variables*, i.e. *given the cross-classification of which is fixed by the product multinomial sampling scheme.* This unique feature of loglinear models and their associated sampling schemes distinguishes them from other forms of linear models. For example, in standard linear model theory with normal error terms one cannot change one set of linear model results into another by conditioning on marginal totals unless one is working with a completely balanced factorial design.

To illustrate these ideas we return to the 2×2×2 table, and the no second-order interaction model with $u_{123(ijk)} = 0$ for i, j, k = 1, 2. For the Poisson or multinomial sampling schemes, the MSS's of Result 1 are the two-dimensional marginal totals, $\{x_{ij+}\}$, $\{x_{i+k}\}$, and $\{x_{+jk}\}$ (except for linearly redundant statistics included for purposes of symmetry). Using Result 2, we have that the MLE's of the $\{m_{ijk}\}$, if they exist, must satisfy the likelihood equations,

$$\begin{aligned} \hat{m}_{ij+} &= x_{ij+}, \qquad i,j = 1,2, \\ \hat{m}_{i+k} &= x_{i+k}, \qquad i,k = 1,2, \\ \hat{m}_{+jk} &= x_{+jk}, \qquad j,k = 1,2. \end{aligned} \tag{3.15}$$

If the second and third dimensions correspond to explanatory variables with $\{x_{+jk}\}$ fixed by design, then we have a product multinomial sampling scheme and the relevant logit model sets $w_{23(jk)} = 0$ for j = 1,2, and k = 1,2. The MSS's are now $\{x_{ij+}\}$ and $\{x_{i+k}\}$ and the likelihood equations are still given by (3.15), since the third set of equations simply represent the

sampling constraints.

D. Computing MLE's for multi-way tables

As we mentioned in Section 2, for many of the special versions of loglinear models such as no-second-order interaction in three-way tables, we need to solve the likelihood equations in expression (3.14) using some type of iterative procedure. The two main competitors are the Iterative Proportional Fitting Procedure (IPFP, e.g. see Bishop, Fienberg, and Holland, 1975), which has linear convergence properties, and Newton's method (or related quadratic convergence algorithms such as the one used in the GLIM package). For a discussion of advantages and disadvantages of each of these methods, and the possibility of using hybrid algorithms, see Fienberg and Meyer (1983).

The IPFP algorithm as implemented in the BMDP package uses a parametrization for loglinear and logit models different from the parametrization in the version of Newton's method used by the GLIM package. Both packages will, however, produce the same estimated expected values satisfying the likelihood equations. What is needed both here, and in the context of linear models more generally, is flexible software that can convert from one parametrization to the other with minimal effort on the part of the user. The technology for doing this already exists. What we need to do as statisticians is remember that the form of parametrization or the choice of a basis for interpreting linear models need not necessarily be the same as the parametrization or basis actually used for doing the computation. All too often we let interpretation drive computation or vice versa. This need not happen.

4. Flexibility of the Loglinear Model Approach

The likelihood results of Section 3C are quite general and apply to large numbers of categorical data problems other than those where the parameters in the model are directly associated with the dimensions of a complete multi-way contingency table. Before the general results had been derived, statisticians had often approached each special problem as a separate enterprise, sometimes using loglinear models and sometimes not. For example, the entire

literature on paired-comparisons (David, 1963), and the Bradley-Terry model (Bradley and Terry, 1952) and its generalizations in particular, was developed without reference to loglinear or logit models per se. Cox (1970) took special note of the logistic form of the Bradley-Terry model in his book on the analysis of binary data, and formal links to the loglinear model theory and literature appeared in Imrey, Johnson, and Koch (1976), Fienberg and Larntz (1976), and Fienberg (1979). Other examples of where categorical data problems have been restructured and analyzed directly using loglinear model techniques include capture-recapture analysis (Fienberg, 1972), latent structure analysis (Goodman, 1974), Guttman scaling (Goodman, 1975), and Milgram's small world problem (Fienberg and Lee, 1975).

Three other topics that have recently been linked to the loglinear model literature are (a) the analysis of censored survival data, (b) the analysis of social and other network data, and (c) the analysis of survey and intelligence test data using the Rasch model. We discuss each in turn, and provide some relevant references.

For the analysis of survival data interest often focuses on the form of the hazard function

$$h(t,x) = f(t|x)/[1 - F(t|x)] \tag{4.1}$$

where $f(t|x)$ and $F(t|x)$ are the pdf and cdf at time t given x, an associated set of fixed covariates. Cox (1972) introduced a proportional hazards model of the form

$$h(t|x) = h_0(t) \cdot e^{x^T \beta} , \tag{4.2}$$

and much of the discussion by Cox and others (such as Breslow in the formal discussion following Cox's paper) made reference to the links between the analysis of expression (4.2) and the categorical data literature (e.g. through Mantel-Haenszel tests). A more formal linkage is possible especially in the case where the covariates, x, are categorical and the underlying hazard function, $h_0(t)$, is piecewise constant (see Holford 1976, 1980). In this case not only is the hazard function loglinear, but so is the likelihood[1] after using a transformation to an "equivalent" Poisson sampling model based on an extension of Birch's (1963) result (Laird and

[1] Actually it is an affine translation of a loglinear model likelihood

Oliver, 1981). The latter authors then show how to estimate β in this special case of Cox's model using IPFP. Related results have been alluded to somewhat less directly by Aitkin and Clayton (1980) and Whitehead (1980) who explain how to use GLIM to estimate censored survival data. An important feature of this result is the ease with which it generalizes to other related survival problems such as those involving competing risks.

A directed graph consists of a set of g nodes, and a collection of directed arcs connecting pairs of nodes. Such graphs have been used to depict social networks describing relationships between pairs of individual actors. Let y be a *sociomatrix* or *adjacency matrix* with elements

$$y_{ij} = \begin{cases} 1 & \text{if a directed arc goes from i to j} \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

where by convention, the diagonal terms $y_{ii} = 0$. Holland and Leinhardt (1981) note that for any pair or *dyad* in a network, with adjacency matrix **y**,

$$y_{ij}y_{ji} + y_{ij}(1-y_{ji}) + (1-y_{ij})y_{ji} + (1-y_{ij})(1-y_{ji}) = 1 , \tag{4.4}$$

for i = j, and that exactly one of the terms on the left hand side of (4.9) is 1 and the remaining three are 0. They then suggest the following model to describe these outcomes (using **Y** as the matrix of random variables of which the adjacency matrix **y** is a realization):

$$\begin{aligned} \log \Pr[(1-Y_{ij})(1-Y_{ji}) = 1] &= \lambda_{ij} \\ \log \Pr[(1-Y_{ij})Y_{ji} = 1] &= \lambda_{ij} + \alpha_j + \beta_i + \theta \\ \log \Pr[Y_{ij}(1-Y_{ji}) = 1] &= \lambda_{ij} + \alpha_i + \beta_j + \theta \\ \log \Pr[Y_{ij}Y_{ji} = 1] &= \lambda_{ij} + \alpha_i + \alpha_j + \beta_i + \beta_j + 2\theta + \rho , \end{aligned} \tag{4.5}$$

where the $\{\lambda_{ij}\}$ are "dyadic" effects included here (but only implicitly in Holland and Leinhardt) to assure that the multinomial constraint (4.4) is satisfied, and where

$$\Sigma_{i=1}^{g} \alpha_i = \Sigma_{j=1}^{g} \beta_j = 0 . \tag{4.6}$$

If we assume that the dyads are independent, then we have a product-multinomial sampling model with one observation per multinomial. Holland and Leinhardt make direct use of

exponential family theory results on maximum likelihood estimation to estimate the parameters in (4.5). Fienberg and Wasserman (1981a, 1981b) note, however, that there is a link between their model and a loglinear model for a multi-dimensional table representation of the probabilities in (4.5). In particular, they work with the four-dimensional array:

$$\begin{aligned} X_{ij11} &= Y_{ij}Y_{ji} \\ X_{ij10} &= Y_{ij}(1-Y_{ji}) \\ X_{ij01} &= (1-Y_{ij})Y_{ji} \\ X_{ij00} &= (1-Y_{ij})(1-Y_{ji}). \end{aligned} \tag{4.7}$$

Note that $X_{ijks} = X_{jisk}$, because the dyad (i,j) is the same as the dyad (j,i). By using this redundant representation, we get a contingency table analogue to the Holland-Leinhardt model. In particular, Meyer (1982) shows that fitting their model via maximum likelihood to $\mathbf{y} = \{y_{ij}\}$ is equivalent to fitting a loglinear model to the newly created redundant array $\{x_{ijks}\}$, i.e. the model of no-second-order interaction.

What is especially attractive about the multi-dimensional contingency table representation of the social network data problem as outlined here is that it generalizes to extensions of the Holland/Leinhardt model (Fienberg and Wasserman, 1981a, 1981b) and it carries over to networks involving multiple relationships. For further details, see Fienberg, Meyer, and Wasserman (1981, 1983).

The final topic of this section also begins with one categorical data representation and ends up with a different but familiar loglinear representation for a multiway table. The results of *ability tests* are often structured in the form of sequences of 1's for correct answers and 0's for incorrect answers. For a test with k problems or items administered to n individuals, we let

$$Y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ answers item } j \text{ correctly} \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Thus we have a two-way table of random variables $\{Y_{ij}\}$ with realizations $\{y_{ij}\}$. An

alternative representation of the data is in the form of a $n \times 2^k$ table $\{W_{ij_1 j_2 \cdots j_k}\}$ where the subscript i still indexes individuals and now $j_1, j_2, \ldots, j_k$ refer to the correctness of the responses on items 1,2,...,k, respectively, i.e.

$$W_{ij_1 j_2 \cdots j_k} = \begin{cases} 1 & \text{if i responds } (j_1, j_2, \ldots j_k) \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

The Rasch model (Rasch, 1960 as reprinted in 1980) for the $\{Y_{ij}\}$ is

$$\log \frac{P(Y_{ij}=1)}{P(Y_{ij}=0)} = \psi + \mu_i + \nu_j , \tag{4.10}$$

where

$$\Sigma \mu_i = \Sigma \nu_j = 0 . \tag{4.11}$$

Expression (4.10) is a *logit* model in the usual contingency table sense for a 3-dimensional array whose first layer is $\{y_{ij}\}$ and whose marginal totals adding across layers is an $n \times k$ table of 1's.

Maximum likelihood estimation for the parameters of the Rasch model (4.10) has been the focus of several authors including Rasch and Andersen. Unconditional maximum likelihood (UML) estimates can be derived but they have rather problematic asymptotic properties, e.g. the estimates are inconsistent as $n \to \infty$ and k remains moderate, although they are consistent when both n and $k \to \infty$ (Haberman, 1977). Fischer (1981) provided an interesting link to the loglinear model literature by approach UML estimation via the embedding of the matrix $y = \{y_{ij}\}$ into a larger $(n+k) \times (n+k)$ matrix of the form:

$$A = \{a_{ij}\} = \begin{bmatrix} 0 & e^T - y^T \\ y & 0 \end{bmatrix} \tag{4.12}$$

where e is an $n \times k$ matrix of 1's. Then he notes that the Rasch model of (4.10) is transformed into an incomplete version of the Bradley-Terry model discussed at the beginning of this section.

Now, we turn to a conditional approach to likelihood estimation (CML) advocated initially by Rasch, who noted that the conditional distribution of Y given the individual marginal totals $\{y_{i+} = y_{i+}\}$ depends only on the item parameters, $\{\nu_j\}$. Then each of the row sums $\{y_{i+}\}$ can take only k+1 distinct values corresponds to the number of correct responses. Next, we recall the alternate representation of the data in the form of an $n \times 2^k$ array, $\{W_{j_1 j_2 \cdots j_k}\}$, as given by expression (4.9). Adding across individuals we create a 2^k contingency table, X, with entries

$$X_{j_1 j_2 \cdots j_k} = W_{\cdot j_1 j_2 \cdots j_k} . \tag{4.13}$$

Duncan (1983) and Tjur (1982) independently noted that we can estimate the item parameters for the Rasch model of (4.10) using the 2^k array x, and the loglinear model

$$\log m_{j_1 j_2 j_3 \cdots j_k} = \omega + \Sigma_{s=1}^{k} \delta_{j_s} \nu_s + \gamma_{j_+} , \tag{4.14}$$

where the subscript $j_+ = \Sigma_{s=1}^{k} j_s$, $\delta_{j_s} = 1$ if $j_s = 1$ and is 0 otherwise, and

$$\Sigma_{r=0}^{k} \gamma_r = 0 . \tag{4.15}$$

The *amazing* result, due to Tjur (1982), is that maximum likelihood estimation of the 2^k contingency table of expected values, $m = \{m_{j_1 j_2 \cdots j_k}\}$ using a Poisson sampling scheme and the loglinear model (4.10), produces the conditional maximum likelihood estimates of $\{\nu_j\}$ for the original Rasch model. Tjur proves this equivalence by (1) assuming that the individual parameters are independent identically distributed random variables from some completely unknown distribution, π; (2) integrating the conditional distribution of Y given $\{Y_{i+}=y_{i+}\}$ over the mixing distribution, π; (3) embedding this "random effects" model in an "extended random model"; and (4) noting that the likelihood for the extended model is equivalent to that for (4.10) applied to x (using Result 3 of Section 3 above). Fienberg (1981) then noted that the model of (4.14) is the model of quasi-symmetry preserving one-dimensional marginal totals, first proposed by Bishop, Fienberg, and Holland (1975, Chapter 8).

Cressie and Holland (1983) have independently developed an approach to the Rasch model similar to Tjur's and Duncan's, and they note other interesting linkages to other aspects of latent trait models.

5. Speculation on Future Developments

The early parts of this paper represent a look backwards upon the development of the methodology for the loglinear model analysis of categorical data. Section 4 is a brief examination of the recent past and the present, stressing how researchers have been effectively adapting the loglinear model approach to new non-standard categorical data problems. In keeping with the spirit of the Conference, I have been allowed, indeed encouraged, to end with some speculation on where we go from here.

The easy part of speculating on the future is to prepare a list of work now underway or work that may begin quite soon:

(a) *Ordinal Variables.* Many papers have been written in recent years extending the comprehensive loglinear approach to problems involving ordinal variables (e.g. Agresti, 1982, or Fienberg, 1982b). A subtle problem described in passing in these papers is the use of monotonicity constraints on loglinear parameters to reflect the ordinal structure. There is also a dual problem involving monotonicity constraints on marginal totals. The computation of MLE's for such models requires attention as does the issue of assessing goodness-of-fit.

(b) *Two Problems in the Analysis of Network Data.* A problem we skimmed by in Section 4 is the lack of relevance of standard asymptotic theory for the loglinear model for network data. The 4-way array used above is of size $4g^2$, with a total count of $2g(g-1)$, while the Holland-Leinhardt model has $2g$ parameters. Haberman (1981) gives some relevant asymptotics for this problem, but more attention is required before the distribution of goodness-of-fit statistics is in hand. A second vexing problem for network data is the assumption of dyadic independence. What is needed is the formulation of a less restrictive model allowing for dyadic dependence, which includes the Holland-Leinhardt model as a special case.

(c) *Logistic Regression Diagnostics.* The logistic regression model is a simple extension of the logit model of Section 3 where the explanatory variables are

continuous rather than categorical (see Fienberg, 1980, Chapter 6). One way to view such models is as corresponding to very sparse cross-classifications, and thus it comes as no surprise that the usual asymptotic theory for overall goodness-of-fit statistics is inapplicable. Landwehr, Pregibon, and Shoemaker (1983) give some interesting graphical devices for logistic regression diagnostics to help with assessing goodness-of-fit. More attention to this problem is needed.

(d) *Computation.* Although computer programs for fitting loglinear and logit models are now widely available, many of the more interesting applications involve very large, sparse arrays that do not fit easily into core in most modern computers. Two directions of research on computations for categorical data will include (1) the development of programs for personal computers that make effective use of disk and auxiliary storage space, and (2) the development of programs for array processors that make effective use of parallel algorithms.

(e) *Bayesian Approaches.* Many papers have been written on Bayesian approaches to the analysis of categorical data using loglinear and logit models. No one has yet to describe an easily implementable Bayesian approach for large multi-way tables.

(f) *Applications.* To date the applications of the loglinear model methodology have occurred primarily in the biological, medical, and social sciences. I believe we can look forward to new applications in agriculture and in industrial settings. In particular, I foresee the use of loglinear model methodology in the development of multivariate quality control techniques.

It is more difficult to look further ahead into the future. Now that we have reached the stage where we have developed a comprehensive approach to categorical data analysis, that parallels the linear model theory for measurement data, I believe we need to step back. As useful as the loglinear model approach has proved to be, it is all too easy to misinterpret loglinear model parameters by imparting inappropriate causal interpretations. I see little hope for a "grand unified theory" applicable to all problems in all settings. The key to understanding longitudinal processes, for example, is the development of formal stochastic

models and their application to observed data. When the data are categorical and are measured at several fixed points in time the issue should be: how well does the underlying stochastic model fit. Instead we tend to fit loglinear or other off-the-shelf models to the resulting cross-classifications, and then to make loose interpretations about the "ideas" in the stochastic model. More careful attention to such problems (e.g. see Cohen and Singer, 1979, Singer and Cohen, 1980, and Singer, 1981) may bear far more interesting results, and will certainly generate difficult statistical problems in need of solution. Thus, for me, the most promising direction for statistical research on categorical data analysis is away from the comprehensive approach described in this paper.

REFERENCES

Agresti, A. (1983). "A survey of strategies for modelling cross-classifications having ordinal variables." *Journal of the American Statistical Association*, 78, 184-189.

Aitkin, M. and Clayton, D. (1980). "The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM." *Applied Statistics*, 29, 156-163.

Andersen, E.B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North Holland.

Baker, R.J. and Nelder, J.A. (1978). *The GLIM System, Release 3, Manual*. Oxford, England: Numerical Algorithms Group.

Bartlett, M.S. (1935). "Contingency table interactions." *Journal of the Royal Statistical Society Supplement*, 2, 248-252.

Birch, M.W. (1963). "Maximum likelihood in three-way contingency tables." *Journal of the Royal Statistical Society B*, 25, 229-233.

Bishop, Y.M.M. (1967). "Multidimensional contingency tables: cell estimates." Ph.D. Dissertation, Department of Statistics, Harvard University. Available from University Microfilm Service.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: MIT Press.

Box, J.F. (1978). *R.A. Fisher, The Life of a Scientist*. New York: Wiley.

Bradley, R.A. and Terry, M.E. (1952). "The rank analysis of incomplete block designs. I. The method of paired comparisons." *Biometrika*, 39, 324-345.

Cochran, W.G. (1940). "The analysis of variance when experimental errors follow the Poisson or binomial laws." *Annals of Mathematical Statistics*, 11, 335-347.

Cochran, W.G. (1942). "The chi-square correction for continuity." *Iowa State College Journal of Science*, 16, 421-436.

Cohen, J.E. and Singer, B. (1979). "Malaria in Nigeria: constrained continuous-time Markov models for discrete-time longitudinal data on human mixed species infections." In *Some Mathematical Questions in Biology* (S. Levin, ed.), Vol. 12. Providence: American Mathematical Society.

Cox, D.R. (1970). *Analysis of Binary Data*. London: Methuen.

Cox, D.R. (1972). "Regression models and life tables" (with discussion). *Journal of the Royal Statistical Society B*, 34, 187-220.

Cressie, N. and Holland, P.W. (1983). "Characterizing the manifest probabilities of latent trait

models." *Psychometrika*, 48, 129-142.

Darroch, J.N. (1962). "Interaction in multi-factor contingency tables." *Journal of the Royal Statistical Society B*, 24, 251-263.

David, H.A. (1963). *The Method of Paired Comparisons*. London: Griffin.

Deming, W.E. and Stephan, F.F. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *Annals of Mathematical Statistics*, 11, 427-444.

Dempster, A.P. (1971). "An overview of multivariate data analysis." *Journal of Multivariate Analysis*, 1, 316-346.

Duncan, O.D. (1983). "Rasch measurement in survey research: further examples and discussion." In *Survey Measurement of Subjective Phenomena* (C.F. Turner and E. Martin (eds.)), Vol. 2 (forthcoming).

Fienberg, S.E. (1972). "The multiple-receptive census for closed populations and incomplete 2^k contingency tables." *Biometrika*, 59, 591-603.

Fienberg, S.E. (1979). "Loglinear representation for paired comparison models with ties and within-pair order effects." *Biometrics*, 35, 479-481.

Fienberg, S.E. (1980). *The Analysis of Cross-classified Categorical Data* (2nd ed.). Cambridge, Mass.: MIT Press.

Fienberg, S.E. (1981). "Recent advances in theory and methods for the analysis of categorical data: making the link to statistical practice." *Bulletin of the International Statistical Institute*, 43rd Session, 49 (Book 2), 763-791.

Fienberg, S.E. (1982a). "Contingency tables." In *Encyclopedia of Statistical Sciences* (S. Kotz and N.L. Johnson, eds.), *Vol.* 2, New York: Wiley, 161-170.

Fienberg, S.E. (1982b). "Using information on ordering for loglinear model analysis of multidimensional contingency tables." *Proc. XIIth International Biometrics Conference*, Toulouse, France, 133-139.

Fienberg, S.E. and Larntz, K. (1976). "Loglinear representation for paired and multiple comparisons models." *Biometrika*, 63, 245-254.

Fienberg, S.E. and Lee, S.K. (1975). "On small world statistics." *Psychometrika*, 40, 219-229.

Fienberg, S.E. and Meyer, M.M. (1983). "Loglinear models and categorical data analysis with psychometric and econometric applications." *Journal of Econometrics* (in press).

Fienberg, S.E., Meyer, M.M., and Wasserman, S.S. (1981). "Analyzing data from multivariate directed graphs: an application to social networks." *Looking at Multivariate Data* (Vic Barnett, ed.), Chichester, England: Wiley, 289-306.

Fienberg, S.E., Meyer, M.M., and Wasserman, S.S. (1983). "Statistical analysis of multiple sociometric relations." Department of Statistics, Carnegie-Mellon University Technical Report No. 245 (revised).

Fienberg, S.E. and Wasserman, S.S. (1981a). "Comment on 'An exponential family of probability distributions for directed graphs'." *Journal of the American Statistical Association*, 76.

Fienberg, S.E. and Wasserman, S.S. (1981b). "Categorical data analysis of single sociometric relations." *Sociological Methodology 1981*, San Francisco: Jossey-Bass, 156-192.

Fischer, G.H. (1981). "On the existence and uniqueness of maximum-likelihood estimates in the Rasch model." *Psychometrika*, 46, 59-78.

Fisher, R.A. (1922). "On the interpretation of χ^2 from contingency tables, and the calculation of p." *Journal of the Royal Statistical Society*, 85, 87-94.

Good, I.J. (1963). "Maximum entropy for hypotheses formulation especially for multi-dimensional contingency tables." *Annals of Mathematical Statistics*, 34, 911-934.

Goodman, L.A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika*, 61, 215-231.

Goodman, L.A. (1975). "A new model for scaling response patterns: an application of the quasi-independence concept." *Journal of the American Statistical Association*, 70, 755-768.

Haberman, S.J. (1974). *The Analysis of Frequency Data*. Chicago: Univ. Chicago Press.

Haberman, S.J. (1977). "Maximum likelihood estimates in exponential response models." *Annals of Statistics*, 5, 815-841.

Haberman, S.J. (1981). "Comment on 'An exponential family of probability distributions for directed graphs.'" *Journal of the American Statistical Association*, 76, 60-61.

Holford, T.R. (1976). "Life tables with concomitant information." *Biometrics*, 32, 587-597.

Holford, T.R. (1980). "The analysis of rates and survivorship using loglinear models." *Biometrics*, 36, 299-306.

Holland, P.W. and Leinhardt, S. (1981). "An exponential family of probability distributions for directed graphs." *Journal of the American Statistical Association*, 76, 33-50.

Imrey, P.B., Johnson, W.D., and Koch, G.G. (1976). "An incomplete contingency table approach to paired comparison experiments." *Journal of the American Statistical Association*, 71, 614-623.

Imrey, P.B., Koch, G.G., Stokes, M.E. in collaboration with Darroch, J.N., and Freeman, D.H., Jr., and Tolley, H.D. (1981). "Categorical data analysis: some reflections on the loglinear model and logistic regression. Part I: Historical and methodological overview." *International Statistical Review*, 49, 265-284.

Imrey, P.B., Koch, G.G., Stokes, M.E. in collaboration with Darroch, J.N., and Freeman, D.H., Jr., and Tolley, H.D. (1981). "Categorical data analysis: some reflections on the loglinear model and logistic regression. Part II: Data analysis." *International Statistical Review*, **50**, 35–64.

Kastenbaum, M.A. (1970). "A review of contingency tables." In *Essays in Probability and Statistics* (S.N. Roy, Chakravarti, P.C. Mahalanobis, C.R. Rao, and H. Smith, eds.). Chapel Hill, N.C.: Univ. North Carolina Press, 407–438.

Kastenbaum, M.A. and Lamphiear, D.E. (1959). "Calculation of chi-square to test the no three-factor interaction hypothesis." *Biometrics*, **15**, 107–115.

Killion, R.A. and Zahn, D.A. (1976). "A bibliography of contingency table literature." *International Statistical Review*, **44**, 71–112.

Laird, N. and Oliver, D. (1981). "Covariance analysis of censored survival data using log-linear analysis techniques." *Journal of the American Statistical Association*, **76**, 231–240.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1983). "Graphical methods for assesssing logistic regression models." *Journal of the American Statistical Association* (forthcoming).

Meyer, M.M. (1982). "Transforming contingency tables." *Annals of Statistics*, **10**, 1172–1181.

Nelder, J.A. and Wedderburn, R.W.M. (1972). "Generalized linear models." *Journal of the Royal Statistical Society A*, **135**, 370–384.

Pearson, K. (1900). "On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling." *Philosophical Magazine, Ser. 5*, **50**, 157–172.

Plackett, R.L. (1983). "Karl Pearson and the chi-squared test." *International Statistical Review*, **51**, 59–72.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* (Reprinting of the 1960 book.) Chicago: Univ. of Chicago Press.

Roy, S.N. and Kastenbaum, M.A. (1956). "On the hypothesis of no 'interaction' in a multiway contingency table." *Annals of Mathematical Statistics*, **27**, 749–757.

Singer, B. (1981). "Estimation of nonstationary Markov chains from panel data." *Sociological Methodology 1981*, San Francisco: Jossey-Bass, 319–337.

Singer, B. and Cohen, J.E. (1980). "Estimating malaria incidence and recovery rates from panel surveys." *Mathematical Biosciences*, **49**, 273–305.

Snedecor, G.W. (1937). *Statistical Methods.* Ames, Iowa: Iowa State Univ. Press.

Snedecor, G.W. (1958). "Chi-square of Bartlett, Mood, and Lancaster in a 2^3 contingency table." *Biometrics*, **14**, 560–562.

Tjur, T. (1982). "A connection between Rasch's item analysis model and a multiplicative Poisson model." *Scandinavian Journal of Statistics*, 9, 23–30.

Whitehead, J. (1980). "Fitting Cox's regression model to survival data using GLIM." *Applied Statistics*, 29, 268–275.

Yule, G.U. (1900). "On the association of attributes in statistics: with illustrations from the material of the childhood society, & c." *Philosophical Transactions of the Royal Society, Ser. A*, 194, 257–319.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 287	2. GOVT ACCESSION NO. AD-A135967
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Towards a Comprehensive Approach to the Analysis of Categorical Data	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Stephen E. Fienberg	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Contracts Office, Carnegie-Mellon University, Pittsburgh, PA 15213	12. REPORT DATE: June, 1983
	13. NUMBER OF PAGES: 27
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Chi-square goodness-of-fit tests; Conditional independence; Cross-classifications; Loglinear models; Interactions; Rectangular arrays

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Current methodology for the analysis of categorical data can be traced to two papers published in 1900, one by Pearson and one by Yule. The keys to the linking of their ideas have been the use of linear models and inferential tool due to Fisher. After 50 years of research, statisticians have come close to developing a comprehensive approach to categorical data problems that stresse three basic themes: interpretability, flexibility, and computability. This paper surveys the evolution of this comprehensive approach and classes of problems for which it has proven useful. A concluding section contains some

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Ent

20. speculation on unsolved methodological problems of current interest and on future developments.

FILM